Scientometrics
https://doi.org/10.1007/s11192-018-2974-5

Two decades of information systems: a bibliometric review

Jiaying Liu[1] · Jiahao Tian[1] · Xiangjie Kong[1] · Ivan Lee[2] · Feng Xia[1]

Received: 11 September 2018

Abstract
Information systems (IS) is a vital research area in both science and engineering, whose revolutions in terms of theories, techniques, and applications promote the evolution of human society. At the same time, the complexity and dynamics of IS raise the challenge for exploring the topic in detail. In this paper, we present a quantitative analysis on bibliographic dataset to reveal the evolution of information systems in 2 decades (1996–2015). We select 39,767 papers published in 8 top-tier journals and conferences between 1996 and 2015 and explore the anatomy from manifold. We find that IS is experiencing the sustainable growth phase in terms of increased productivity, impact, and collaboration. The field is benefited from collaborative, open-minded, and in-depth efforts evidenced from the growing number of co-authors per paper, the continual declining of self-citation rate, and increased reference age, respectively. By applying topic detection models on paper titles, abstracts, and keywords, we infer the representative topics and research directions, which can also reveal the research landscape within this field. Finally, we measure the temporal trends of topics and identify the innovative years in the 20 years' development history of IS. These discoveries can benefit not only researchers in terms of promoting understanding of the entire field, but also governments for funding agencies.

Keywords Information systems · Statistical analysis · Data analytics · Science of science

Introduction

The development of information systems (IS) has undertaken a unique growth trajectory in the recent past, as it encompasses various topics including computer networking, database management, decision support systems, information security, as well as system analysis and design. The early phase of IS has been committed to the study of business models and related algorithmic processes, that are used to build the IT systems to bridge various disciplines of Computer Science and Business. Subsequently, IS has been evolved to cover various related

✉ Feng Xia
f.xia@ieee.org

[1] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China

[2] School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia

Published online: 15 December 2018

disciplines such as Computer Science, Engineering, Mathematics, Management Science, Cybernetics. As IS has been undertaking rapid development, there's a need to review the background of IS and investigate how it evolves over time.

The importance of reviewing a discipline's current state is self-evident. For example, it helps newcomers understanding the key developments in the field, and it also helps institutions setting relevant policies to define strategic directions to follow. A good way to reflect the development trend of a specific research field is to analyze the scientific publication, which is often considered as a key proxy in both theory and practice. Until now, a great body of literature has been devoted to studying the publication data from the perspective of network modeling and analysis using quantitative methods in Scientometrics. Correia et al. (2018) provide a quantitative assessment of mapping the intellectual structure of CSCW and its development. Sinatra et al. (2015) reveal the rapid growth and inner structure of Physics based on the Web of Science data spanning more than 100 years. Similarly, Liu et al. (2018) study the evolution of Artificial Intelligence relying on the ability of the complex network topologies. Although these studies highlight the evolution as well as the problems and challenges changing over time for the specific field, they fail to provide topic-specific information for the researchers to better understand the changing research context over time. This paper aims to provide an overview of the IS research and profile the field by studying its main trends from 1996 to 2015. To fill the gaps mentioned above, we are particularly interested in gaining insight into the following questions:

- How does IS change over time in terms of the volume of relevant literature? Does the trend in IS increase or decrease? Does IS attract growing number of researchers?
- What is the pattern of citation fluctuation over the past years? Does a generic pattern exist for researchers' reference behaviour?
- What are the milestones in the 20-year development of IS? How to identify the impactful papers, researchers, and institutions as well as their unique characteristics?
- How can we explore the inner structure of IS researches?
- What topics are of concern to the authors in IS? What is their relatedness with respect to their history and future progress? How are they interacting with each other?
- How can we quantify the importance of each year to profile the field from the perspective of temporal variations? How to identify crucial years in which the key changes occurred?

The remainder of the paper is structured as follows. "Methodology" section summarizes methodologies used in this study, including the introduction of the dataset collection and preprocessing. We also present in detail the measures used to quantify the development of IS at levels of growth of the publication, topics detection and evolution, and important years identification. "Results and analysis" section provides the overview of our findings from the perspective of the growth of the domain, impacts and its variations, the inner structure of IS, and the future of the domain. Finally, "Conclusion" section concludes the study, summarizes limitations, and suggests potential future directions.

Methodology

Dataset

The publication metadata is obtained from Microsoft Academic Graph (MAG),[1] which is a widely used database and also one of the best-curated databases for empirical research in scientometrics and citation analysis (Sinha et al. 2015). We select the articles from 8 conferences and 8 journals in the field of IS research from 1996 to 2015. The list of top-tier journals is obtained in the Association for Information Systems(AIS).[2] The list of top-tier conferences relevant to "Information Systems" is obtained from the Computing Research and Education Association of Australasia (CORE).[3] Specifically, we search for "A*" ranked conferences listed in "CORE2018" with their "Field of Research" equals "0806".[4]

The analysis in this paper considers different attributes between journal and conference publication. Generally, conferences are held every one or two years, providing opportunities for scholars to exchange and discuss research findings. Conference publications have relatively short review cycle (typically 2–4 months), which is essential for topics that require timely dissemination. In contrast, journal papers are usually presents a comprehensive study on a research topic, taking a relative long review cycle with multiple rounds of review. These differences lead to the fluctuation in publications/citations growth among research topics, resulting in different popularities of hot topics for the area.

Tables 1 and 2 present the summary statistics for the selected journals and conferences. Assume that the set of selected publication metadata for the specific journal/conference is P, using elementary statistics including total number of publications $|P|$, total number of authors $\sum_{p\in P}|\mathrm{au}_p|$, and total number of citations $\sum_{p\in P}|\mathrm{ci}_p|$ from 1996 to 2015, we calculate the average number of citations per paper $\sum_{p\in P}|\mathrm{ci}_p|/|P|$, the average number of authors per paper $\sum_{p\in P}|\mathrm{au}_p|/|P|$, and the average number of paper per author $|P|/\sum_{p\in P}|\mathrm{au}_p|$ in the period of study.

Furthermore, to reflect the dynamics in researchers' reference behavior, we adopt the most rigorous definition of self-citation as our evaluation altimetric, that is, if the citing paper and the cited paper have at least one common author, then the citation between these two papers is an author self-citation. Using consistent criterion on journals and conferences, the self-journal/conference citation is that two papers are published on the same journal/conference.

Topic detection and evolution

Topic detection

A major defect of the publication datasets is that there is not a certain topic classification for each paper. To quantify the variation of topics across journals and conferences, we use *Latent Dirichlet Allocation (LDA)* (Blei et al. 2003) for topic modeling. LDA is an unsupervised machine learning technique built on the classical probabilistic latent

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.
[2] http://aisnet.org/?SeniorScholarBasket.
[3] http://www.core.edu.au.
[4] http://portal.core.edu.au/conf-ranks/?search=0806&by=for&source=CORE2018&sort=arank&page=1.

Table 1 Journal statistics

Journal name	Papers	Authors	Citations	Citations per paper	Authors per paper	Papers per author
European Journal of Information Systems	941	1933	31050	33.00	2.05	0.68
Information Systems	414	1008	19872	48.00	2.43	0.54
Information Systems Research	702	1903	70291	100.13	2.71	0.64
Journal of Information Technology	564	1159	17141	30.39	2.05	0.67
Journal of Management Information Systems	818	2214	66525	81.33	2.71	0.61
Journal of the Association for Information Systems	338	826	11275	33.36	2.44	0.52
MIS Quarterly	832	2034	148856	178.91	2.44	0.71
The Journal of Strategic Information Systems	425	952	20320	47.81	2.24	0.60

Table 2 Conference statistics

Title	Acronym	Papers	Authors	Citations	Citations per paper	Authors per paper	Papers per author
ACM Conference on Human Factors in Computing Systems	CHI	11340	37598	240651	21.22	3.32	0.66
ACM Conference on Economics and Computation	EC	1464	3964	52535	35.88	2.71	0.61
International Conference on Information Systems	ICIS	6743	18702	28280	4.19	2.77	0.60
IEEE Information Visualization Conference	InfoVis	316	890	12761	40.38	2.82	0.51
IEEE International Symposium on Wearable Computing	ISWC	4437	15338	68664	15.48	3.46	0.45
Joint Conference on Digital Libraries	JCDL	1369	4414	13359	9.76	3.22	0.52
ACM International Conference on Research and Development in Information Retrieval	SIGIR	3718	10979	131121	35.27	2.95	0.79
International World Wide Web Conference	WWW	5346	17378	206723	38.67	3.25	0.51

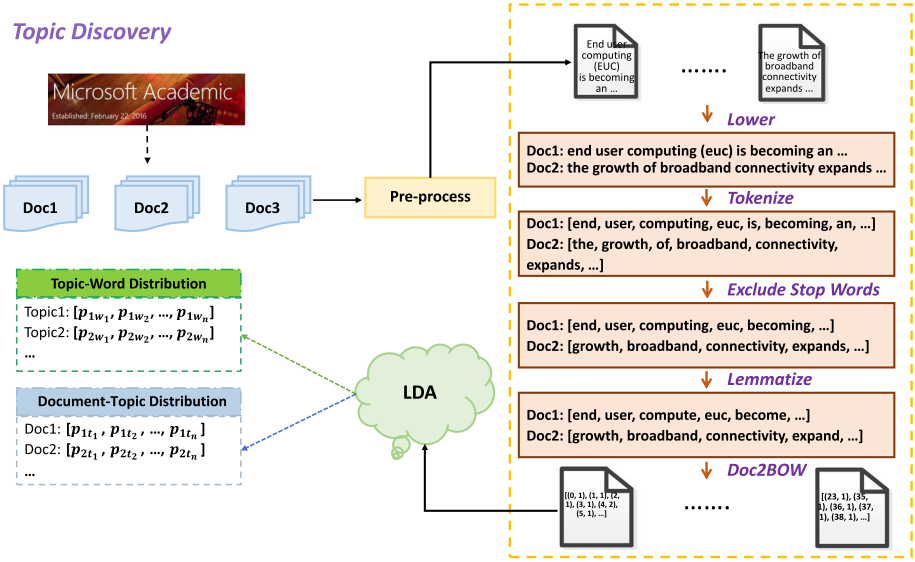


Fig. 1 Graphical model representation of topic modeling process

semantic analysis (pLSA) model (Hofmann 2017) that can be applied in discovering main themes from the large-scale document-word corpus. As a three-layer Bayesian Probability model, which is consisted of three layers of words, topics, and documents, it does not require any label of documents and can form topics naturally from the statistical structure of document-word data itself. For a corpus D consisting of M documents each of length N_i, the generative process can be summarized into three steps:

1. Use $\theta_t \sim \text{Dirichlet}(\alpha)$ to find the topic distribution for each topic t
2. Use $\psi_d \sim \text{Dirichlet}(\beta)$ to find the topic distribution for each document d
3. For each word $w_{i,j}$, where $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, N_i\}$:
   - Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_t)$
   - Choose a word $w_{i,j} \sim \text{Multinomial}(\psi_{z_{i,j}})$

Figure 1 shows the graphical representation of our topic modeling process in plate notation. We use the title, keywords, and the abstract as the input of LDA model. In order to get more accurate results, we first pre-process the data according to following steps:

1. Convert all uppercase initials to lowercase.
2. Segment sentences into independent words.
3. Remove stopwords, i.e. common but topic-irrelevant words like is, are, have, in the context.
4. Lemmatize words so that words appearing in inflected forms (e.g. computing and compute, firms and firm) can be identified as same words.
5. Convert documents into bags-of-words representation.

Table 3 Notations of variables and parameters

Notation	Description
y	Exact publication year for paper p
$\hat{y}_p$	Predicted publication year for paper p
Y_s	Starting point of the study period (i.e. 1997)
Y_e	End point of the study period (i.e. 2016)
P_y	The set of papers published in year y
Past_y	The set of papers published in y satisfying the condition: $\hat{y}_p < y$
Future_y	The set of papers published in y satisfying the condition: $\hat{y}_p > y$

After data preprocessing, the metadata in terms of the bag of words are used as the input of LDA model. Finally, we can get topic-word distribution for each topic and topic-document distribution for each paper.

Topic evolution

Our purpose is not only confined to topic detection within the field. We are also interested in the temporal variation for hot topics over the study period. In order to reveal the overall trend from the perspective of time dimension, we adopt following indicators to evaluate topics.

Topic distribution over time. How to capture the dynamic of the specific topic within the field over time? We use similar measures as the work in Sun and Yin (2017) and use θ_t^k to represent the proportion of the topic k at year t. Furthermore, to investigate the topics, we divide the study period into two time periods 1996–2005 and 2006–2015, of which contain 10 years. Then the increase index for each topic is defined as $II_k = \Sigma_{t=2006}^{2015}\theta_t^k / \Sigma_{t=1996}^{2005}\theta_t^k$.

Turnaround years identification

Following Wang's step (Savov et al. 2017), in order to identify turnaround years, the first step is to identify the publication year for each paper. If its topic distribution matches the topics distribution of papers published in the near future, then we consider this paper as an innovative paper. The importance of each year is determined by the number of innovative papers published in that year.

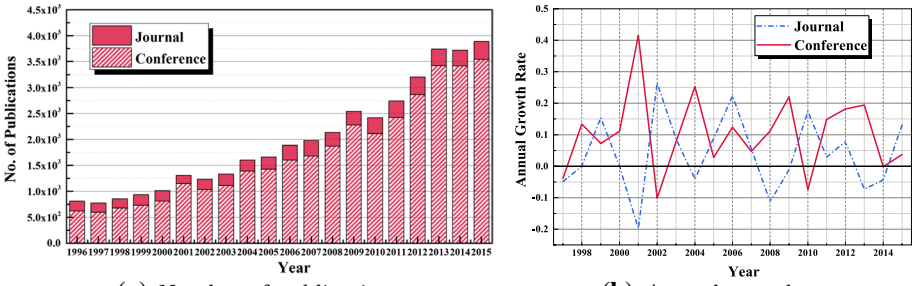
To illustrate the particular process of turnaround years identification, we list the notations of variables and parameters in Table 3.

The innovation score for each year can be computed as:

$$S(y) = \text{ERR}_{\text{F}_y}/|\text{P}_y| \cdot N_{\text{F}_y} - \text{ERR}_{\text{P}_y}/|\text{P}_y| \cdot N_{\text{P}_y} \tag{1}$$

where ERR_{F_y} and ERR_{P_y} represent the total prediction error for all papers in Future_y and Past_y, respectively. In order to eliminate the bias of y's position, we regard N_{F_y} and N_{P_y} as the normalization factors for future papers and past papers.

After the computation, years with relatively high innovation score are considered as turnaround years.



(a) Number of publications **(b)** Annual growth rate

Fig. 2 The evolution of IS in terms of publications' growth. **a** Yearly IS publications from 1996 to 2015. **b** Annual growth rate of the publications

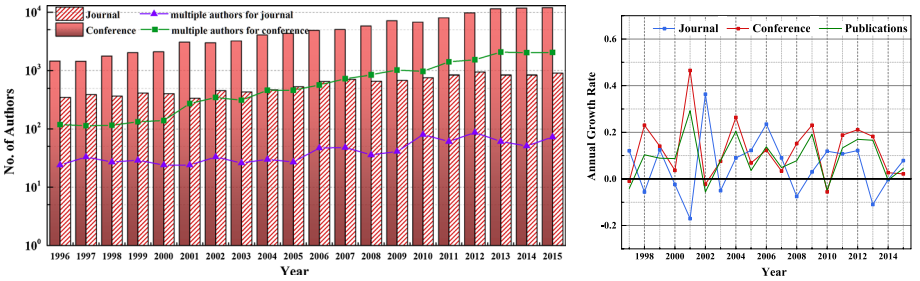
Results and analysis

In this section, we focus on showing and interpreting out main results according to following aspects which can answer the questions mentioned in "Introduction" section: (1) Try to find out whether IS is growing from perspectives of publication volume and author volume. (2) Analyze the impact in terms of citation pattern analysis and capture scholars' reference behavior dynamics. (3) Discover important milestones (i.e. important papers, scholars, institutions in this field). (4) Detect topics variation over time and their inner connections. (5) Identify the turnaround year.

The growth of IS

Figure 2a presents the yearly number of publications during the studying period. As a consequence, the overall volume of publications for both selected journals and conferences has achieved sustained growth. The increasing trend can also be reflected by the linear fitting parameters (Table 4). At the same time, we discover that the number of journal publications is far less than conference publications. We also perform an inspection on the publication growth rate (Fig. 2b) across years in detail. Although results of publication volume reflect a slow but sustainable ascending tendency, indeed, we can discover short-term fluctuations in their growth rate clearly.

The growth results may be driven by the quantity and quality of authors, acceptance rate, as well as some internal policies. From the perspective of the quantity of authors (Fig. 3), we find that the growth rate of scholars follow the same trend as the academic papers. On the one hand, the phenomenon may be due to the continued increase in the number of papers. On the other hand, the increasing number of co-authors for a publication is another potential cause. In order to emphasize the growth on IS, we use the size of publications and authors to measure the scholars' average productivity and collaboration patterns (as shown in Fig. 4). The comparison of the two results reveals that the scientific collaboration is becoming more common which can be evidenced from the increasing average number of co-authors per paper and the individual productivity has dropped by 15% from 1996 to 2015.

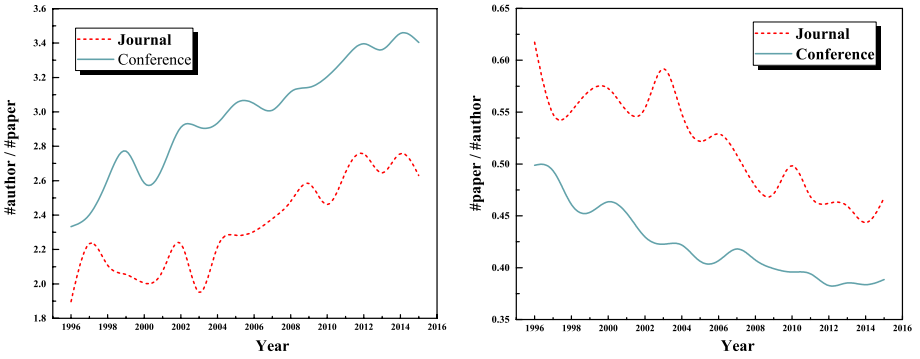


(a) Number of authors **(b)** Annual growth rate

Fig. 3 The evolution of IS from the perspective of the number of authors. **a** The yearly number of authors who have published papers in the selected journals and conferences. **b** The annual growth rate of authors and overall publications

Table 4 Linear fitting details for publication volume

x	y	Fitted equation	Pearsons' r	R^2
Year	No. of journal papers	$y = 9.65x - 19174.76$	0.93	0.87
Year	No. of conference papers	$y = 159x - 317979.25$	0.96	0.93

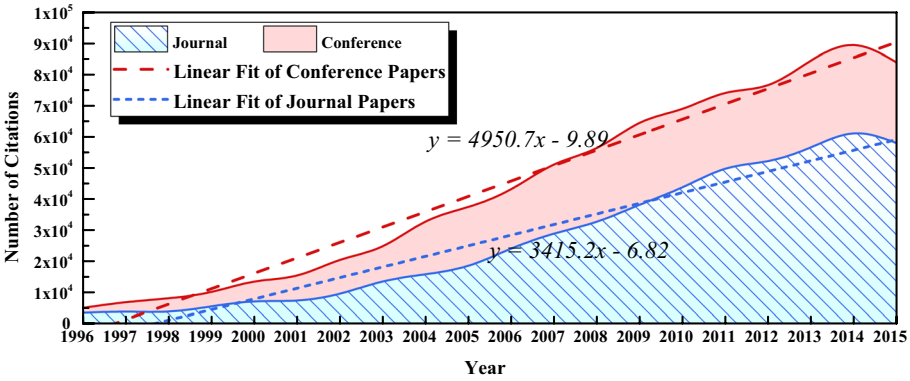


(a) Number of authors in a paper **(b)** Authors' average productivity

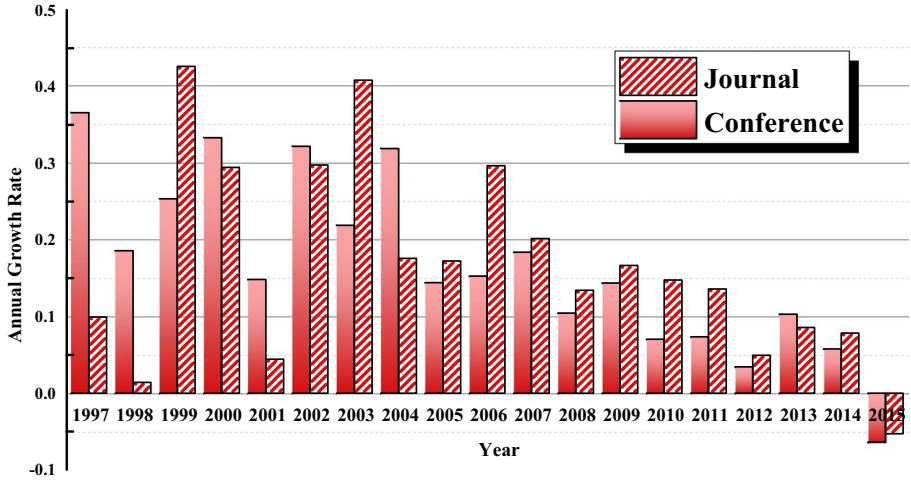
Fig. 4 The evolution of IS from the perspectives of collaboration patterns and average productivity. **a** The yearly number of authors in a paper which is calculated as the rate between the overall size of publications and authors. **b** Authors' productivity considers the unique number of authors

IS impact analysis

Science is developed upon knowledge discovery. It can be regarded as a gift economy (Bollen et al. 2009) and the value of itself can be measured by the contribution to the new acknowledgment or ideas of others. Scholars cite related literature to announce that the new work they have established is inspired by the prior work. Therefore researchers usually use citations a paper received to represent its impact. In this paper, we use collective citations of selected IS papers to analyze the overall trend of this area.



(a) Number of citations over time



(b) Annual growth rate of citations

Fig. 5 The changes of IS impact over time. **a** The yearly total number of citations. **b** The annual growth rate of citations

In Fig. 5a, there is a clear increasing trend of the total number of citations over time. The growth rate was nearly 40% at the beginning of the study period and dropped to 10% later. Although the number of new citations each year has been increasing, the growth rate has slowed down. The drastic increase actually exists in all fields partly due to the exponential growth of literature. The development of modern science is another possible reason.

The citing behavior has been changing within the field of IS over the study period. To capture the citing behavior dynamics, we seek to analyze the impact from the perspective of individuals. Here we regard a citation as a self-citation if the citing paper and the cited paper have at least one common author. It reflects a scholar's tendency to cite their own papers. Similar to the definition of author-self citation, we calculate the self-citation rate for each journal and conference. That is, if the cited paper published in the same journal or conference with the paper that cites it, then the citation is a journal (or conference) self-citation.

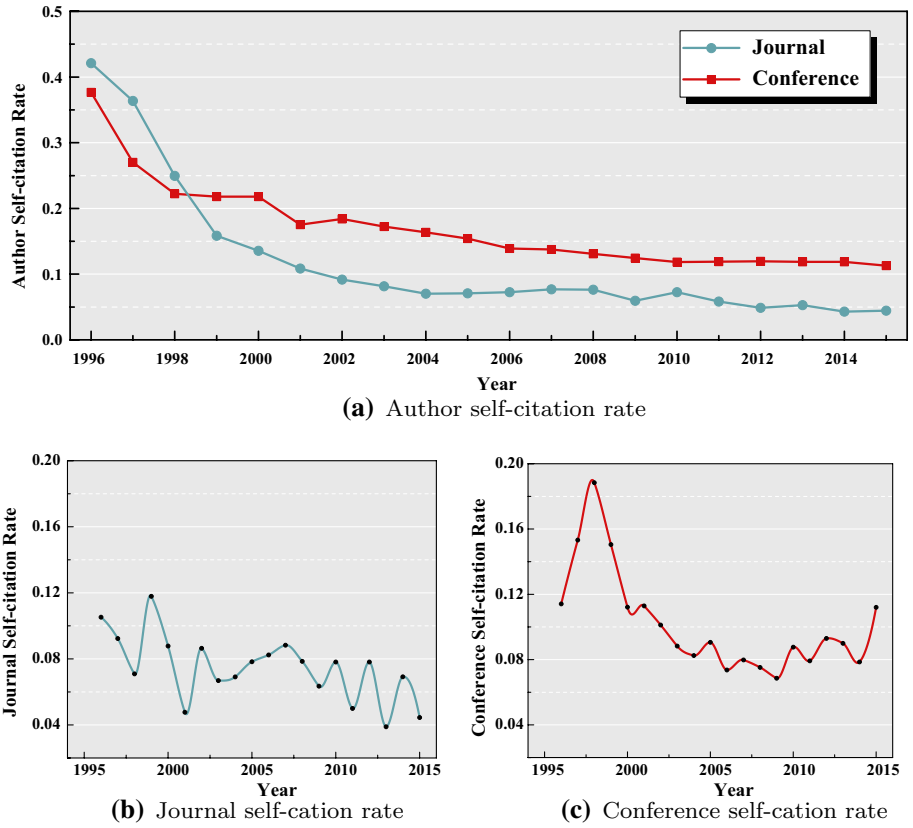
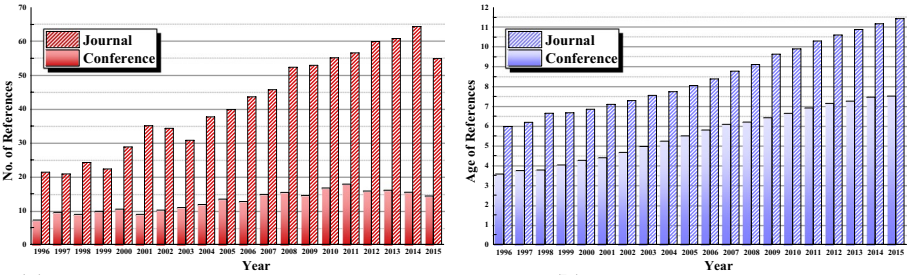


Fig. 6 The citing behavior evolution from the perspectives of three types of self-citation. **a** The yearly author self-citation rate. An author-self citation means that paper and its citation have at least one common author. **b** The journal self-citation indicates that two papers are published in the same journal. **c** The conference self-citation is the citation which comes from the same conference paper

Figure 6 shows the results of self-citation analysis. Authors' tendency to cite their own papers has been falling down from 40% in 1996 to only 8–10% in 2015. In fact, the journal self-citation rate and the conference self-citation rate behave quite same as the author-self citation rate does (see Fig. 6b, c). In the context of self-citing behavior, the decreasing trend unveils that this area is becoming more and more open-minded.

Finally, we aim to uncover the relationship between sustained growth of citations and a paper's reference list size. The most likely reason for the explosion of citations is the increase in the size of the reference list in a paper over time. We are also interested in the age which can be defined as the year difference between a paper's published time and the published year of papers that cite it. It can reveal authors' behaviors that how many years they look back in the literature. Figure 7a shows the average number of references in different years. The drastic increase in the size of reference lists is in part because there are more and more open academic search engines and databases with the development of modern science which make scholars become more accessible to the literature. The reason for the great difference in reference number between journal papers and conference papers is that they vary greatly in focus, scope, and length. Generally, the scope and the focus of



(a) Average number of references over time **(b)** Average age of references

Fig. 7 The evolution of reference behavior. **a** The yearly average size of the reference list in a paper. **b** The age of the reference is defined as the time between the publish year of a paper and the year of papers that cite it

journal papers are larger than conference ones. Survey papers which will cite more papers are often published in journals rather than conferences because conference papers usually have limitations of space.

In the study period, a gradual increase trend appears in the average reference age as shown in Fig. 7b. Scholars' tendency to cite "old" papers is increasingly evident, which is evidenced by nearly 100% growth during 2 decades. Taking journal papers as an example, the average reference age of them stays at 6 years in 1996 and reaches nearly 11.5 years in 2015. The evolution of science makes it much easier for scholars to search and access old publications.

Influential identities identification in IS

Tables 5 and 6 present the top 30 cited journal papers and conference papers, respectively. Concerning the total citation count in the study period, three survey papers have received the most number of citations (for journal papers).

Actually, the total number of citations is cumulative over time, the results only show the temporary ranking because earlier published papers have cumulative advantages. In the field of scientometrics, various metrics to evaluate papers are proposed (Bai et al. 2017; Haslam et al. 2008), while there's lack of a universal standard to quantify the prestige level of papers.

To evaluate the contribution of each IS author, we study the distribution of citations per author. Table 7 presents the top cited authors. Some authors are prolific because the collaboration becomes more common. So we regard the average number of citations per paper as the indicator to evaluate scholars in this field instead of the number of papers.

Inner structure of IS

Following steps we have mentioned in "Topic detection" section, we obtain the posterior topic distribution of each paper and posterior word distribution of each topic. Actually, deciding the number of topics for LDA model is one of the greatest challenges within the field. There has been an increasing interest in scientific topics detection (Griffiths and Steyvers 2004; Cao et al. 2009; Arun et al. 2010) which is a hot issue within science. The topics we decided are affected by the data statistical structure. In

Table 5 Top 30 cited journal papers in 1996–2015

No.	Title	Citations	Year
1	User acceptance of information technology: toward a unified view	8682	2003
2	Review: knowledge management and knowledge management systems: conceptual foundations and research issues	4689	2001
3	Design science in information systems research	4499	2004
4	The DeLone and McLean model of information systems success: a 10-year update	3472	2003
5	Trust and TAM in online shopping: an integrated model	2853	2003
6	A set of principles for conducting and evaluating interpretive field studies in information systems	2824	1999
7	Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model	2341	2000
8	A partial least squares latent variable modeling approach for measuring interaction effects: results from a monte carlo simulation study and an electronic-mail emotion/adoption study	2308	2003
9	A resource-based perspective on information technology capability and firm performance: An empirical investigation	2097	2000
10	Understanding information systems continuance: an expectation-confirmation model	2044	2001
11	Issues and opinion on structural equation modeling	1974	1998
12	Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior	1951	2000
13	Developing and validating trust measures for e-Commerce: an integrative typology	1945	2002
14	Why should I share? Examining social capital and knowledge contribution in electronic networks of practice	1939	2005
15	Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs	1858	1999
16	Time flies when you're having fun: cognitive absorption and beliefs about information technology usage	1796	2000
17	Beyond accuracy: what data quality means to data consumers	1654	1996
18	Knowledge management: an organizational capabilities perspective	1573	2001
19	Analyzing the past to prepare for the future: writing a literature review	1537	2002
20	Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social-psychological factors, and organizational climate	1472	2005
21	Qualitative research in information systems	1457	1997
22	Review: information technology and organizational performance: an integrative model of it business value	1454	2004
23	Applying the technology acceptance model and flow theory to online consumer behavior	1424	2002
24	Research commentary: desperately seeking the IT in IT research—a call to theorizing the IT artifact	1379	2001

Table 5 (continued)

No.	Title	Citations	Year
25	Social cognitive theory and individual reactions to computing technology: a longitudinal study	1260	1999
26	User acceptance of hedonic information systems	1253	2004
27	Gender differences in the perception and use of E-mail: an extension to the technology acceptance model	1247	1997
28	A respecification and extension of the DeLone and McLean model of IS success	1227	1997
29	Review: the resource-based view and information systems research: review, extension, and suggestions for future research	1226	2004
30	A conceptual and operational definition of personal innovativeness in the domain of information technology	1165	1998

Table 6 Top 30 cited conference papers in 1996–2015

No.	Title	Citations	Year
1	The anatomy of a large-scale hypertextual Web search engine	9251	1998
2	Item-based collaborative filtering recommendation algorithms	3246	2001
3	Understanding and using context	2933	2001
4	Towards a better understanding of context and context-awareness	2705	1999
5	Probabilistic latent semantic indexing	2660	1999
6	Tangible bits: towards seamless interfaces between people, bits and atoms	2423	1997
7	Comparison of multiobjective evolutionary algorithms: empirical results	2422	2000
8	What is Twitter, a social network or a news media?	2398	2010
9	The Eigentrust algorithm for reputation management in P2P networks	2364	2003
10	Talking about tactile experiences	2228	2013
11	A re-examination of text categorization methods	1898	1999
12	A language modeling approach to information retrieval	1888	1998
13	An algorithmic framework for performing collaborative filtering	1759	1999
14	Contextual design: defining customer-centered systems	1727	1997
15	Earthquake shakes Twitter users: real-time event detection by social sensors	1479	2010
16	Reality mining: sensing complex social systems	1476	2006
17	Uniform Resource Identifiers (URI): generic syntax	1429	1998
18	A study of smoothing methods for language models applied to Ad Hoc information retrieval	1413	2001
19	Stanford encyclopedia of philosophy	1373	2011
20	Labeling images with a computer game	1293	2004
21	The use of MMR, diversity-based reranking for reordering documents and producing summaries	1267	1998
22	A taxonomy of web search	1218	2002
23	A survey on context-aware systems	1197	2007
24	Approximating the nondominated front using the Pareto archived evolution strategy	1177	2000
25	Focused crawling: a new approach to topic-specific Web resource discovery	1171	1999

Table 6 (continued)

No.	Title	Citations	Year
26	Analysis of recommendation algorithms for e-commerce	1158	2000
27	Extensible Markup Language (XML)	1150	1997
28	Mining the peanut gallery: opinion extraction and semantic classification of product reviews	1129	2003
29	Yago: a core of semantic knowledge	1078	2007
30	Electronic commerce	1071	2008

Table 7 Top 30 cited authors in 1996–2015

No.	Name	Organization	Citations	#Papers	Citations per Paper
1	Sergey Brin	Stanford University	9251	1	9251
2	Fred D. Davis	University of Arkansas	8682	1	8682
3	Lawrence Page	Stanford University	9939	2	4969.5
4	Dorothy E. Leidner	Texas Christian University	4689	1	4689
5	Salvatore T. March	Own Graduate School of Management	4499	1	4499
6	Alan R. Hevner	College of Business Administration	4499	1	4499
7	Sudha Ram	Eller College of Business And Public Administration	4499	1	4499
8	Jinsoo Park	College of Business Administration	4499	1	4499
9	Michael G. Morris	University of Virginia	8790	2	4395
10	Ephraim R. McLean	Information Technology Department Kogod School of Business	3472	1	3472
11	William H. DeLone	Information Technology Department Kogod School of Business	3472	1	3472
12	Viswanath Venkatesh	University of Maryland	9417	3	3139
13	Anind K. Dey	College of Computing and Gvu Center	2933	1	2933
14	Peter J. Brown	The University of Kent at Canterbury	2705	1	2705
15	Pete Steggles	Atandt Laboratories Cambridge	2705	1	2705
16	Mark E. Smith	Hewlett Packard Laboratories	2705	1	2705
17	Thomas Hofmann	International Computer Science Institute	2660	1	2660
18	Lothar Thiele	Department of Electrical Engineering	2422	1	2422
19	Eckart Zitzler	Department of Electrical Engineering	2422	1	2422
20	Mario T. Schlosser	Stanford University	2364	1	2364
21	Barbara L. Marcolin	University of Calgary	2308	1	2308
22	Wynne W. Chin	University of Calgary	2308	1	2308
23	Peter R. Newsted	University of Calgary	2308	1	2308
24	George Karypis	Grouplens Research Group Army Hpc Research Center	4404	2	2202
25	John Riedl	Grouplens Research Group Army Hpc Research Center	4404	2	2202

Table 7 (continued)

No.	Name	Organization	Citations	#Papers	Citations per Paper
26	Badrul M. Sarwar	Grouplens Research Group Army Hpc Research Center	4404	2	2202
27	Joseph A. Konstan	Grouplens Research Group Army Hpc Research Center	4404	2	2202
28	Anol Bhattacherjee	University of South Florida	2044	1	2044
29	Charles J. Kacmar	Michigan State University	1945	1	1945
30	Vivek Choudhury	Michigan State University	1945	1	1945

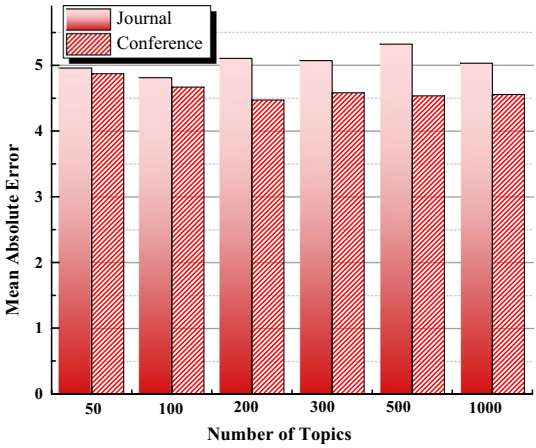
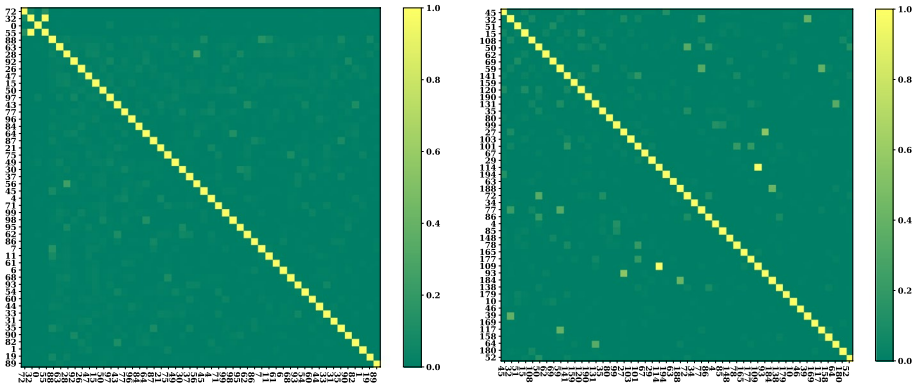


Fig. 8 *Mean Absolute Error (MAE)* for the different number of topics



(a) Co-occurrence probability for hot topics in journals

(b) Co-occurrence probability for hot topics in conferences

Fig. 9 Normalized co-occurrence probability for hot topics in the study period

this paper, we aim to analyze the overall development of IS from a global perspective, so we only use the basic LDA model and make decisions based on results of Support Vector Machine (SVM) (Hearst et al. 1998), which is used to identify the important year of IS. To measure the accuracy of our prediction model, we compute the *Mean Absolute Error (MAE)* as $E_p = \hat{y}_p - y_p$ over all folds in a fivefold cross-validation. Figure 8 presents MAE in different topic numbers (i.e. 50, 100, 200, 300, 500, and 1000).

As balanced with experience and intuition, we finally decide the number of topics: 100 topics for journal papers and 200 topics for conference papers. Tables 8 and 9 present top 50 topics with the highest posterior probability in terms of word clouds. The size of each word is presented proportionately according to its probability. To explore the interaction among these topics, we calculate the co-occurrence probability, that is, the probability of two topics appearing in the same paper and visualize the results as the heat map (see Fig. 9).

Table 8 Word-cloud of popular topics of journal papers

#72 Information Management System	#32 Information Management System	#0 Information Acquisition	#55 Information Management System	#88 Research Methodology
				
#63 Organization Management	#28 Information System Development	#92 Knowledge Management	#26 Technology Acceptance Model	#47 Information Technology Adoption
				
#15 Organization Management	#50 Decision Support System	#97 Firm Management	#43 Electronic Commerce	#77 Social Network
				
#96 ERP	#84 Job Satisfaction	#64 Virtual Team	#87 Structural Equation Model	#21 Electronic Commerce
				
#75 Transaction Cost Theory	#49 Business Information Technology	#30 Network Analysis	#37 Service Management	#56 Risk Management
				
#45 Social Technical System	#4 Online Auction	#71 Information Technology Investment	#99 Human Computer Interaction	#98 Group Communication
				
#95 Open Source Software	#62 Information Market	#86 Behavior Analysis	#7 Research Framework	#11 Assimilation Theory
				
#61 Computer Science	#6 Supply Chain Management	#68 Economic Information	#93 Outsource	#54 Financial Performance
				
#60 Public Sector Information	#44 Health Care	#33 Interorganizational Relationship	#31 Global Digital Divide	#35 Construct Validity
#90 Online Community	#82 Participation Theory	#1 Analytical Model	#19 Information System Design	#89 Artificial Intelligence

From the results we can see that the topics clustered by the LDA model correspond to the research areas. For example, the most popular journal topic is #72: "information, system, technology, management..." which corresponds to the paper category "Information Management System" in the MAG dataset. The most popular conference topic #45: "design, technology, home, work..." corresponds to "Home Technology". We also discover that popular topics differ among journal and conference papers, with journal papers more concentrate on "Information Acquisition" and "Information Management", whereas conference papers pay more attentions to "User Interface Design" and "Entity Extraction".

Table 9 Word-cloud of popular topics of conference papers

#45 Home Technology	#32 Information Retrieval System	#51 User Interface Design	#15 User Center Design	#108 Entity Extraction
				
#50 Organization Theory	#62 User Interface	#69 Interaction Techniques	#59 Search Engine	#141 Hci Research
				
#159 Semantic Web	#120 Research Framework Design	#190 Personal Information Management	#131 Web Browser	#35 Project Management
				
#80 Collaborative System	#99 Digital Library	#27 Prediction Method	#103 Evaluation Method	#101 IS Research
				
#67 Social Network	#29 Context Aware System	#114 Genetic Algorithm	#194 Health Care	#63 Large Scale Data
				
#188 Wireless Network	#72 Business Process Management	#34 Visualization	#77 Search Term Suggestion	#86 Firm Performance
				
#4 Open Source Software	#85 Educational Technology	#148 Mobile Device	#78 Computer Human Interaction	#165 Online Auction
				
#177 World Wide Web	#109 Optimization Problem	#93 Recommendation System	#184 Wireless Sensor Network	#138 Electronic Commerce
				
#179 Game Design	#10 Display Screen	#46 GPS	#39 Probabilistic Model	#169 User Analysis
#117 User Behavior	#158 Online Advertisement	#64 Internet Application	#180 Real Time Information	#52 Knowledge Management

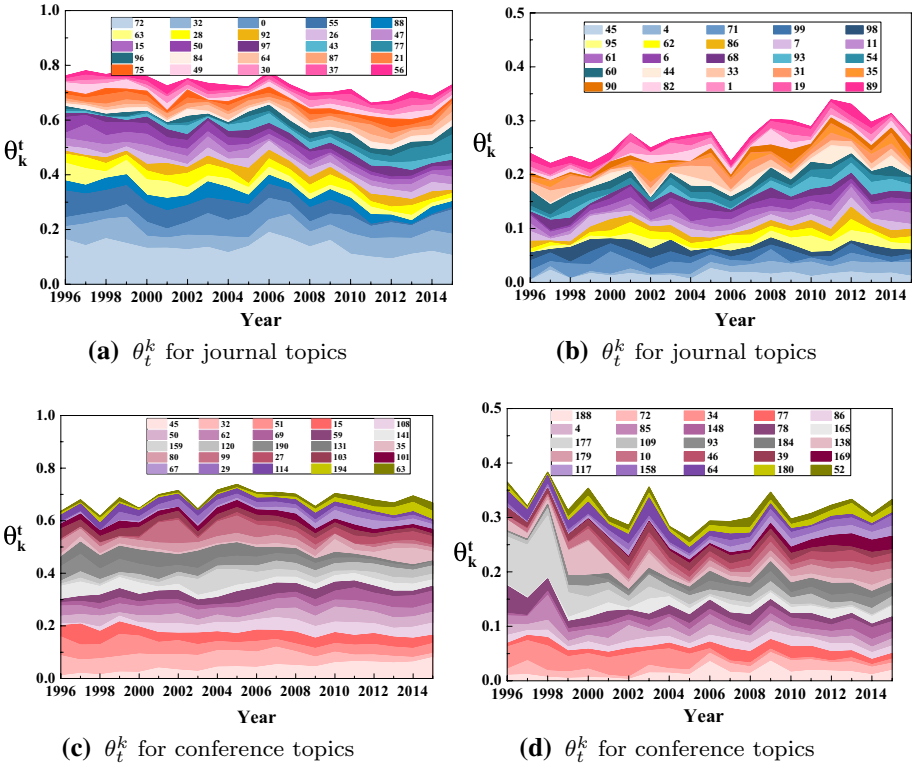
It is important to note that the topic discovery process may be reflected by the following aspects: (1) the total number of papers of a sub-field. (2) the variation of topics for different journals and conferences. The topics may change over time and could not be distinguished using only the overall topic distribution. So after the topic detection process, we calculate θ_t^k for each topic to help uncover the correlations between topics and time variable. In this sense, we focus on capturing the temporal dynamics of each hot topic and measuring the overall development of IS in the studying period. As shown in Fig. 10, we can find not only the proportion of the popular topics in the descending order, but also the change tendencies of these topics. The fluctuation of the line reflects the popularity of the topic over time.

Table 10 Increasing index for hot topics

Journal			Conference		
Topic	Occurrence	II	Topic	Occurrence	II
72	1067	0.92	45	1472	2.10
32	511	1.21	32	1209	0.60
0	434	1.10	51	1163	0.64
55	339	0.39	15	1140	0.90
88	265	0.68	108	1122	1.76
63	235	0.48	50	1097	1.48
28	232	0.81	62	1014	1.00
92	204	1.03	69	904	1.94
26	194	1.13	59	819	1.04
47	192	0.69	141	797	0.76
15	191	0.65	159	774	0.77
50	190	0.42	120	766	2.01
97	168	1.19	190	759	0.52
43	160	1.62	131	737	0.33
77	152	3.82	35	699	2.93
96	140	2.39	80	690	0.74
84	137	1.34	99	666	0.51
64	137	1.93	27	589	5.72
87	135	1.35	103	565	0.66
21	133	0.86	101	533	0.83
75	133	1.21	67	519	4.56
49	130	0.91	29	507	0.94
30	126	1.38	114	501	0.63
37	126	1.43	194	499	3.13
56	125	0.84	63	493	1.82
45	122	1.20	188	466	2.39
4	107	4.03	72	426	1.37
71	106	0.47	34	412	0.17
99	103	0.74	77	409	1.68
98	98	0.67	86	407	1.23
95	97	2.87	4	397	0.86
62	96	1.22	85	392	0.67
86	96	1.15	148	392	1.97
7	96	0.97	78	387	0.57
11	95	1.88	165	385	1.34
61	94	1.01	177	385	0.10
6	92	1.05	109	380	0.90
68	92	1.18	93	371	1.80
93	91	2.87	184	369	3.65
54	91	0.97	138	363	0.41
60	90	0.79	179	361	4.11
44	85	1.76	10	349	0.85
33	85	0.71	46	342	2.43
31	81	1.46	39	342	0.76

Table 10 (continued)

Journal			Conference		
Topic	Occurrence	II	Topic	Occurrence	II
35	76	1.12	169	340	8.52
90	74	2.55	117	328	2.85
82	73	1.34	158	315	2.19
1	71	0.88	64	306	0.37
19	71	1.33	180	305	1.68
89	70	0.87	52	296	1.24



(a) θ_t^k for journal topics **(b)** θ_t^k for journal topics

(c) θ_t^k for conference topics **(d)** θ_t^k for conference topics

Fig. 10 θ_t^k for each topic to help uncover the correlations between topics and time variable

We can observe that some topics have been increasing over time such as topic #89, #90 for journal papers. At the same time, some topics are declining, e.g., topic #109 for conference papers.

Furthermore, to investigate the overall trend of these topics more clearly, the Increase Index II_k for each hot topic k are computed (see Table 10). When $II_k > 1$, it indicates that the topic becomes hotter in the latest 10 years, while $II_k < 1$ suggests the hotness of the

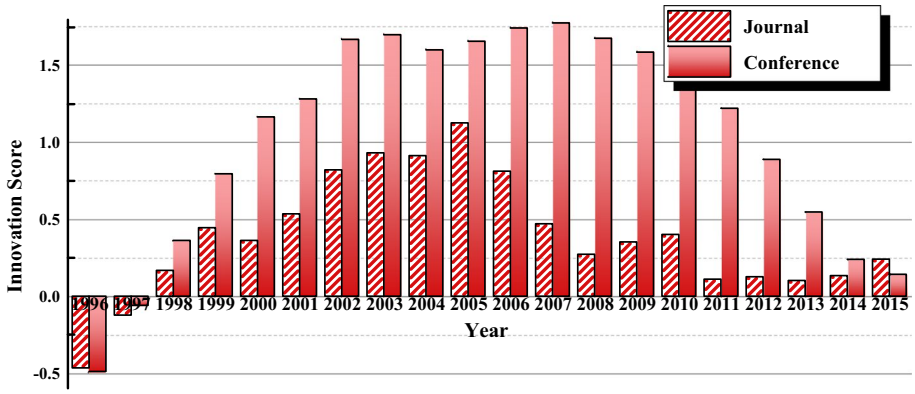
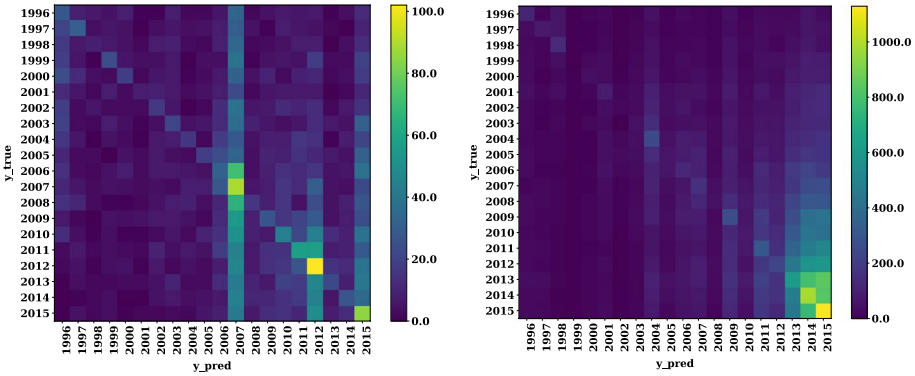


Fig. 11 Innovation score S_y of each year



(a) Confusion matrix for journal hot topics

(b) Confusion matrix for conference hot topics

Fig. 12 Confusion matrix

topic has a decreasing trend over time. From this table, we can easily find the recent development of the topics. For example, for the conference papers, although topic #32 occurs many times in the papers, it becomes less popular comparing with previous 10 years. On the contrary, topic #184 is an emerging topic evidenced by its high *II*.

Turnaround year identification

To study how IS develops over the course of time, we identify the key years when research breakthroughs happen. Based on the classification approach mentioned in "Turnaround years identification" section, first of all, we need to predict the publication dates of all papers. Here we regard all topics distribution as a whole. The method can measure not only the importance of topics over time, but also the relationships among topics' probabilities.

Figure 11 presents $S(y)$ which can represent the importance of each year. In the field of IS, 2003, 2004, 2005 could be considered as turnaround years according to our definition.

We use the heat map to show the confusion matrix, where $x[y_{\text{true}}][y_{\text{pred}}]$ means the number of papers which published in year y_{true} but predicts as year y_{pred}. In Fig. 12, darker color of the elements represent smaller numbers. The concentration of mistakes occurs in 2012 for journal hot topics and 2013–2015 for conference hot topics. The phenomenon indicates that papers are not easy to be distinguished within these periods depending on their topic distributions.

Conclusion

In this paper, we explore the anatomy of IS spanning from 1996 to 2015. The scientometric exercise is carried out on the large bibliographic data provided by MAG to investigate answers for questions mentioned in "Introduction" section. By allowing the assessment of a large number of papers, we unveil the evolutionary patterns along the lines of growth, collaboration, impact, and focus topics in this field. With respect to the questions, the findings are concluded as follows:

- IS publications have increased over the study period, which reflects its rapid development in the recent past. At the same time, the increasing number of researchers also shows the growing popularity of this field. A gradual sustained process of growth in the number of co-authors in a paper indicates that the IS is moving from individual work to collaboration.
- We observe that the influence of IS continues to expand over the study period. The decrease in the self-citation rate by year illustrates that the focus of researchers transforms from their own efforts to others' efforts, which ensures the evolving nature of IS.
- The influential identities we identified in terms of papers and authors undoubtedly emerge as vital milestones with their distinctive contributions to the rapid growth in IS.
- Given the results of the topic detection and analysis, some topics remain consistent over the study period. These topics have not developed independently but interrelated. The interrelationships of all these topics make IS a complex system.
- By distinguishing innovation years in which key changes occurred, we give the importance score of each year in the development of IS. The signposted research breakthroughs of IS appeared in 2003, 2004, and 2005 make them the turnaround years of IS.

In summary, these results suggest that IS, an interdisciplinary and complex field, benefits from the continuous growth of collaboration, openness, and diversity. These results contribute to promoting our understanding of nature and development patterns of this field. By unveiling the hidden insights behind the existing knowledge, this study is meant to serve as a guideline for technical innovations, subject-specific assessments, and funding policies.

Despite our manifold analysis of IS, there are still several perspectives for future directions. First of all, some other techniques within Network Science, for example, citation weighting assessments (Yan and Ding 2010; Zhu et al. 2015; Bai et al. 2018) can be valuable for recognizing the influential entities. Another line of potential research is to adopt other in-depth topic models, such as author topic model (Ngo et al. 2016; Yang

and Hsu 2016) and dynamic topic models (Xu et al. 2017; Finin et al. 2016) to find more insights into understanding the research trends. In addition, it would be important to explore the interplay between the development of IS and the advances in other fields.

Author's Contribution JL, FX, IL, and XK designed the research; JT performed the experiments; XK and FX analyzed the data; JL and IL wrote the paper. All authors reviewed the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402). Springer.

Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., et al. (2017). An overview on evaluating and predicting scholarly article impact. *Information*, *8*(3), 73.

Bai, X., Zhang, F., Hou, J., Lee, I., Kong, X., Tolba, A., et al. (2018). Quantifying the impact of scholarly papers based on higher-order weighted citations. *PLoS ONE*, *13*(3), e0193192.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, *4*(6), e6022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781.

Correia, A., Paredes, H., & Fonseca, B. (2018). Scientometric analysis of scientific publications in CSCW. *Scientometrics*, *114*(1), 31–89.

Finin, T., Cane, M., Sleeman, J., Halem, M., et al. (2016). Dynamic topic modeling to infer the influence of research citations on IPCC assessment reports. In *Big data challenges, research, and technologies in the earth and planetary sciences workshop, IEEE int. conf. on big data*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., et al. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics*, *76*(1), 169–185.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28.

Hofmann, T. (2017). Probabilistic latent semantic indexing. In *ACM SIGIR forum* (Vol. 51, pp. 211–218). ACM.

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., et al. (2018). Artificial intelligence in the 21st century. *IEEE Access*, *6*, 34403–34421.

Ngo, G. H., Eickhoff, S. B., Fox, P. T., & Yeo, B. T. (2016). Collapsed variational bayesian inference of the author-topic model: Application to large-scale coordinate-based meta-analysis. In *2016 international workshop on pattern recognition in neuroimaging (PRNI)* (pp. 1–4). IEEE.

Savov, P., Jatowt, A., & Nielek, R. (2017). Towards understanding the evolution of the WWW conference. In *Proceedings of the 26th international conference on world wide web companion* (pp. 835–836). International World Wide Web Conferences Steering Committee.

Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A.-L. (2015). A century of physics. *Nature Physics*, *11*(10), 791.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). An overview of microsoft academic service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246). ACM.

Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C Emerging Technologies*, *77*, 49–66.

Xu, Z., Chen, L., Dai, Y., & Chen, G. (2017). A dynamic topic model and matrix factorization-based travel recommendation method exploiting ubiquitous data. *IEEE Transactions on Multimedia*, *19*(8), 1933–1945.

Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the Association for Information Science and Technology*, *61*(8), 1635–1643.

Yang, M., & Hsu, W. H. (2016). Hdpauthor: A new hybrid author-topic model using latent dirichlet allocation and hierarchical dirichlet processes. In *Proceedings of the 25th international conference companion on world wide web* (pp. 619–624). International World Wide Web Conferences Steering Committee.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427.